price="8500.0"/>

<car make="Toyota" model="4Runner" year="1990" price="7500.0"/>

From the foregoing description, it should be apparent that the present invention provides a technique for extracting data from structured documents.

Although the invention has been described in detail with reference only to the presently preferred hardware environment, those of ordinary skill in the art will appreciate that various modifications can be made without departing from the invention. Accordingly, the invention is defined only by the following claims.

WHAT IS CLAIMED IS:

1. A method for extracting records from a structured text in a computer system, comprising:

identifying potential locations of values of record fields in the text by identifying locations in the text of items in lists of known potential values for record fields,

identifying a region of interest in the text by applying multiple candidate region partitioners, evaluating each to measure how well it isolates a region with a high density and a high amount of potential locations of values of record fields, selecting one that measures best, and applying it to produce a region of interest,

segmenting the region of interest into record regions that each contain data for a single record by applying multiple candidate segmenters, evaluating each to measure how well it segments into regions such that each region has one field value per record field and such that different regions have similar numbers of field values for each record field, selecting one that measures best, applying it to produce record regions,

extracting field values from record regions by identifying most likely locations of field values for each record field in each record region, and

outputting records composed of extracted field values for record fields.

2. The method of claim 1, with the addition of:

identifying potential locations of values of record fields in the text by identifying locations in the text of patterns of potential values for record fields.

3. The method of claim 1, with the addition of:

identifying potential locations of values of record fields in the text by identifying locations in the text of numbers in ranges that are potential values for record fields.

4. An apparatus for extracting data from a file, comprising a computer and a computer program, performed by the computer, for:

identifying potential locations of values of record fields in the text by identifying locations in the text of items in lists of known potential values for record fields,

identifying a region of interest in the text by applying multiple candidate region partitioners, evaluating each to measure how well it isolates a region with a high density and a high amount of potential locations of values of record fields, selecting one that measures best, and applying it to produce a region of interest,

segmenting the region of interest into record regions that each contain data for a single record by applying multiple candidate segmenters, evaluating each to measure how well it segments into regions such that each region has one field value per record field and such that different regions have similar numbers of field values for each record field, selecting one that measures best, applying it to produce record regions,

extracting field values from record regions by identifying most likely locations of field values for each record field in each record region, and

outputting records composed of extracted field values for record fields.